SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution

Coimbatore-35

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

19ECB212 – DIGITAL SIGNAL PROCESSING

II YEAR/ IV SEMESTER

UNIT 4 – FINITE WORD LENGTH EFFECTS

TOPIC – **TRUNCATION AND ROUNDING**

QUANTIZATION BY TRUNCATION AND ROUNDING

- In fixed point or floating point arithmetic the size of the result of an operation (Sum or Product) may be exceeding the size of binary used in the number system. In such cases the low order bits has to be eliminated in order to store the result
- The two methods of eliminating these low order bits are **Truncation and Rounding.** This process is also referred to as quantization via truncation and rounding
- The effect of rounding and truncation is to introduce an error whose value depends on the number of bits eliminated.

QUANTIZATION STEPS

- The decimal numbers that are encountered as filter coefficients, Sum, Product, etc.... in DSP applications will lie in the range -1 to +1
- When “B” bit binary is selected to represent the decimal numbers, then 2^B binary codes are possible. Hence the range of decimal numbers has to be divided in to 2^B steps and each step is represented by a binary code. Each step of decimal number is also called quantization step

$$\text{Quantization step size, } q = \frac{R}{2^B} = \frac{1-(-1)}{2^B} = \frac{2}{2^B} = \frac{1}{2^B - 2^{-1}}$$

$$= \frac{1}{2^{B-1}} = \frac{1}{2^b} = 2^{-b}$$

Where, R = Range of decimal number

B = Size of binary including sign bit

b = B – 1 = Size of binary excluding sign bit

QUANTIZATION STEPS FOR B=3 AND B-1=2

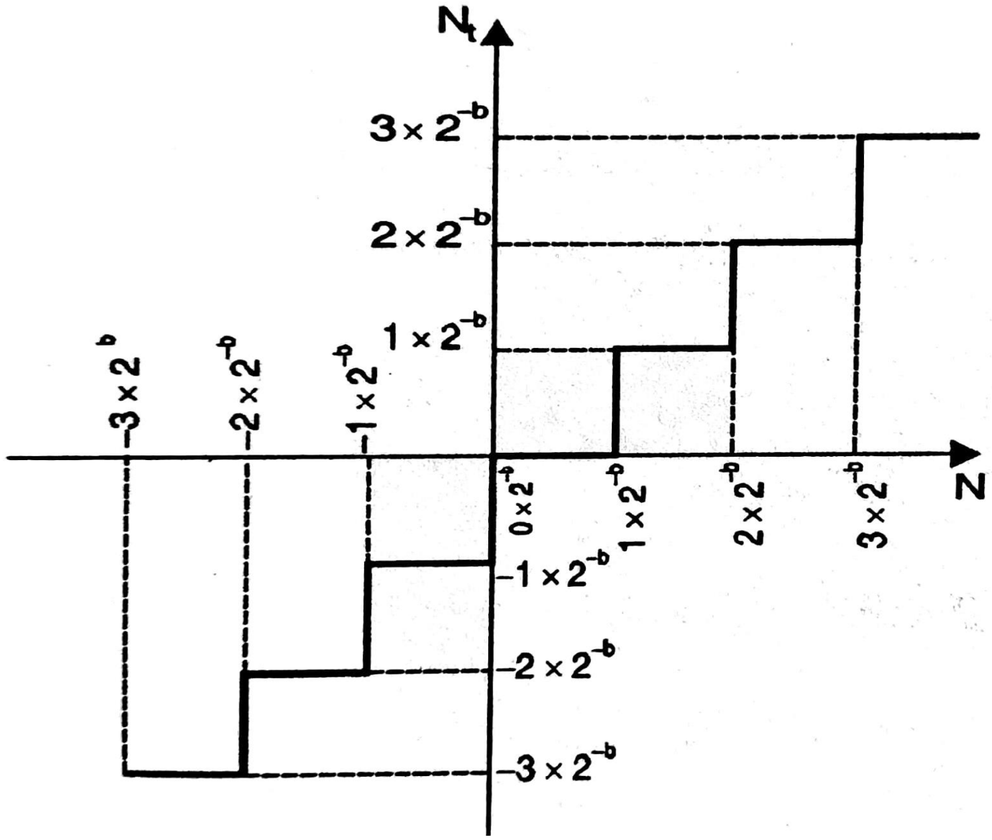
Binary Code	Quantization Steps		
	Sign-magnitude	One's complement	Two's complement
000	$+0 \times 2^{-2} = +0 \times \frac{1}{4} = +0$	$+0 \times 2^{-2} = +0 \times \frac{1}{4} = +0$	$+0 \times 2^{-2} = +0 \times \frac{1}{4} = +0$
001	$+1 \times 2^{-2} = +1 \times \frac{1}{4} = +0.25$	$+1 \times 2^{-2} = +1 \times \frac{1}{4} = +0.25$	$+1 \times 2^{-2} = +1 \times \frac{1}{4} = +0.25$
010	$+2 \times 2^{-2} = +2 \times \frac{1}{4} = +0.50$	$+2 \times 2^{-2} = +2 \times \frac{1}{4} = +0.50$	$+2 \times 2^{-2} = +2 \times \frac{1}{4} = +0.50$
011	$+3 \times 2^{-2} = +3 \times \frac{1}{4} = +0.75$	$+3 \times 2^{-2} = +3 \times \frac{1}{4} = +0.75$	$+3 \times 2^{-2} = +3 \times \frac{1}{4} = +0.75$
100	$-0 \times 2^{-2} = -0 \times \frac{1}{4} = -0$	$-3 \times 2^{-2} = -3 \times \frac{1}{4} = -0.75$	$-4 \times 2^{-2} = -4 \times \frac{1}{4} = -1.00$
101	$-1 \times 2^{-2} = -1 \times \frac{1}{4} = -0.25$	$-2 \times 2^{-2} = -2 \times \frac{1}{4} = -0.50$	$-3 \times 2^{-2} = -3 \times \frac{1}{4} = -0.75$
110	$-2 \times 2^{-2} = -2 \times \frac{1}{4} = -0.50$	$-1 \times 2^{-2} = -1 \times \frac{1}{4} = -0.25$	$-2 \times 2^{-2} = -2 \times \frac{1}{4} = -0.50$
111	$-3 \times 2^{-2} = -3 \times \frac{1}{4} = -0.75$	$-0 \times 2^{-2} = -0 \times \frac{1}{4} = -0$	$-1 \times 2^{-2} = -1 \times \frac{1}{4} = -0.25$

TRUNCATION

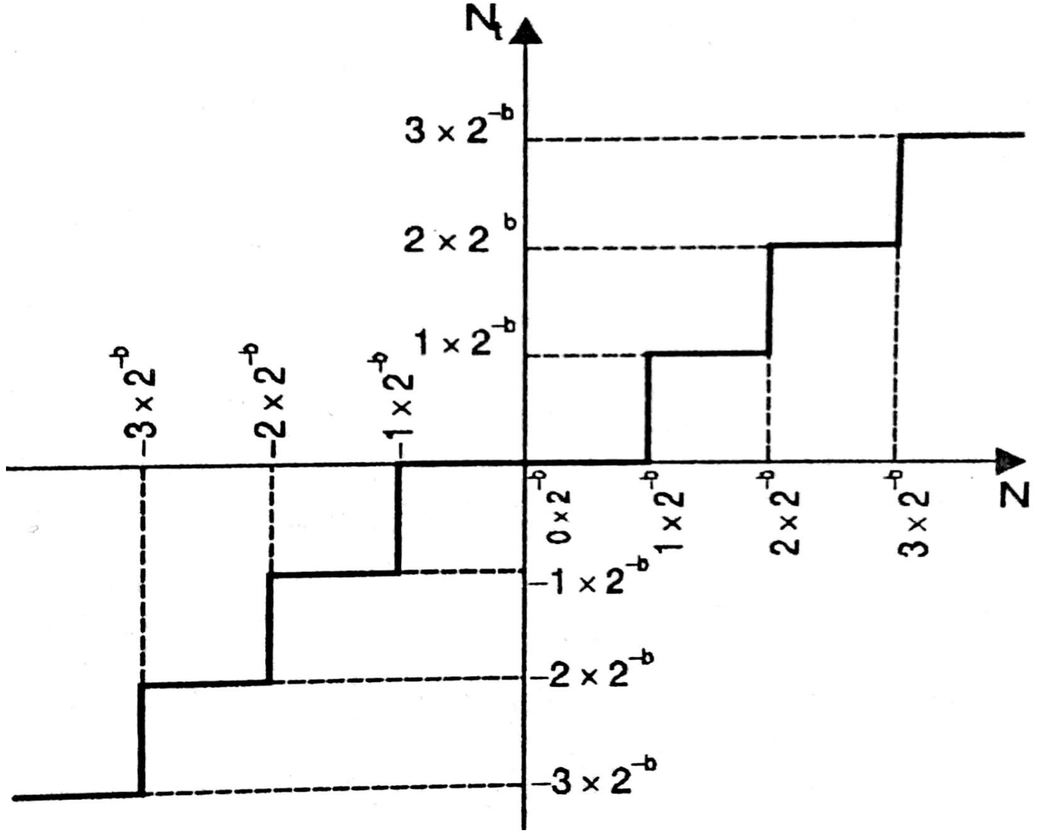
- The truncation is the process of reducing the size of binary number (or reducing the number of bits in a binary number) by discarding all bits less significant than the least significant bit that is retained
- In the truncation of a binary number to b bits, all the less significant bits beyond b^{th} bit are discarded
- In fixed point number system there are three different types of number representation. The effect of truncation on positive numbers are same in all the three representations (because the format for positive number is same in all the three representations)

TRUNCATION CHARACTERISTICS OF QUANTIZER

Two's Complement Quantizer



One's Complement Quantizer



THE INPUT – OUTPUT CHARACTERISTICS OF THE QUANTIZER USED FOR TRUNCATION

1. Any positive unquantized number in the range, $0 \leq N < (1 \times 2^{-b})$, will be assigned the quantization step, 0×2^{-b}.

2. Any positive unquantized number in the range, $(1 \times 2^{-b}) \leq N < (2 \times 2^{-b})$, will be assigned the quantization step, 1×2^{-b} and so on.

3. In sign-magnitude and one's complement quantizer, any negative unquantized number in the range, $(-1 \times 2^{-b}) < N \leq 0$, will be assigned the quantization step, 0×2^{-b}.

THE INPUT – OUTPUT CHARACTERISTICS OF THE QUANTIZER USED FOR TRUNCATION

4. In sign-magnitude and one's complement quantizer, any negative unquantized number in the range, $(-2 \times 2^{-b}) < N \leq (-1 \times 2^{-b})$, will be assigned the quantization step, -1×2^{-b} and so on.

5. In two's complement quantizer, any negative unquantized number in the range, $(-1 \times 2^{-b}) \leq N < 0$, will be assigned the quantization step, -1×2^{-b}.

6. In two's complement quantizer, any negative unquantized number in the range, $(-2 \times 2^{-b}) \leq N < (-1 \times 2^{-b})$, will be assigned the quantization step, -2×2^{-b} and so on.

TRUNCATION

- The error due to truncation of negative number depends on the type of representation of the number
- Let N = Unquantized fixed point binary number
- N_t = Fixed point binary number quantized by truncation
- The quantization error in fixed point number due to truncation is defined a

$$\textbf{Truncation Error } e_t = N_t - N$$

- The truncation of a positive number results in a number that is smaller than the unquantized number. Hence the truncation error is always negative when positive number is trunacted

CASE 1 – POSITIVE NUMBER

The unquantized positive number in the range,

$$(1\times 2^{-b}) \le N < (2\times 2^{-b}) \xrightarrow[\text{truncated to}]{\text{is}} N_t = 1\times 2^{-b}$$

$$\therefore \text{ Minimum error } = 1\times 2^{-b} - 2\times 2^{-b} = -2^{-b}$$

$$\text{Maximum error } = 1\times 2^{-b} - 1\times 2^{-b} = 0$$

$$\therefore \text{ Range of error } = -2^{-b} < e \le 0$$

CASE 2 – SIGN MAGNITUDE AND ONE'S COMPLEMENT NEGATIVE NUMBER

The unquantized negative number in the range,

$$(-2\times 2^{-b}) < N \le (-1\times 2^{-b}) \xrightarrow[\text{truncated to}]{\text{is}} N_t = -1\times 2^{-b}$$

$$\therefore \text{ Minimum error } = -1\times 2^{-b} - (-1\times 2^{-b}) = 0$$

$$\text{Maximum error } = -1\times 2^{-b} - (-2\times 2^{-b}) = 2^{-b}$$

$$\therefore \text{ Range of error } = 0 \le e < 2^{-b}$$

CASE 3 – TWO'S COMPLEMENT NEGATIVE NUMBER

The unquantized negative number in the range,

$$(-1\times 2^{-b}) < N \le (-2\times 2^{-b}) \xrightarrow[\text{truncated to}]{\text{is}} N_t = -2\times 2^{-b}$$

$$\therefore \text{ Minimum error } = -2\times 2^{-b} - (-1\times 2^{-b}) = -2^{-b}$$

$$\text{Maximum error } = -2\times 2^{-b} - (-2\times 2^{-b}) = 0$$

$$\therefore \text{ Range of error } = -2^{-b} < e \le 0$$

RANGE OF ERRORS IN TRUNCATION OF FIXED POINT NUMBERS

Number and its representation	Range of error when truncated to b bits
Positive number	$-2^{-b} < e \leq 0$
Sign - magnitude negative number	$0 \leq e < 2^{-b}$
One's complement negative number	$0 \leq e < 2^{-b}$
Two's complement negative number	$-2^{-b} < e \leq 0$

TRUNCATION

- For the truncation of negative numbers represented in sign magnitude and one's complement format the error is always positive because the truncation basically reduces the magnitude of the numbers
- In the two's complement representation, the effect of truncation on a negative number is to increase the magnitude of the negative number and so the truncation error is always negative
- In floating point representation the mantissa of the number alone is truncated. The truncated error in a floating point number is proportional to the number being quantized

TRUNCATION

- Let N_f = Unquantized floating point binary number
- N_{tf} = Truncated floating point binary number

$$N_{tf} = N_f + N_f\ \varepsilon_t$$

- Where ε_t - relative error due to truncation of a floating point binary number
- Relative error due to truncation is

$$\varepsilon_t = N_{tf} - N_f\ /\ N_f$$

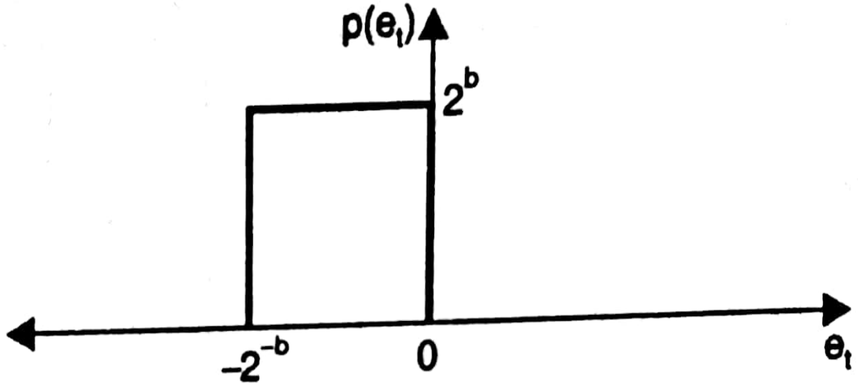
- In truncation of binary number the range of error is known but the probability of obtaining an error within the range is not known

RANGE OF ERRORS IN TRUNCATION OF FLOATING POINT NUMBERS

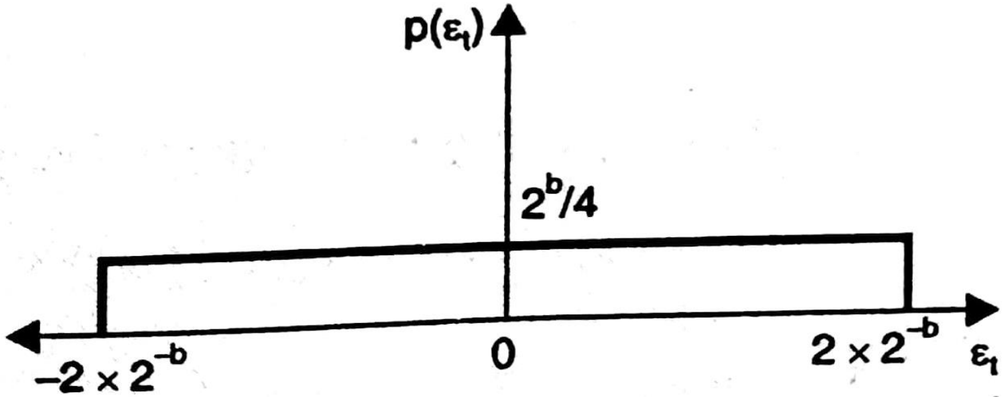
Type of representation for mantissa	Range of error when mantissa is truncated to b bits
Two's complement positive mantissa	$-2 \times 2^{-b} < \varepsilon_t \leq 0$
Two's complement negative mantissa	$0 \leq \varepsilon_t < 2^{-b} \times 2$
One's complement positive and negative mantissa	$-2 \times 2^{-b} < \varepsilon_t \leq 0$
Sign-magnitude positive and negative mantissa	$-2 \times 2^{-b} < \varepsilon_t \leq 0$

QUANTIZATION NOISE PROBABILITY DENSITY FUNCTIONS FOR TRUNCATION

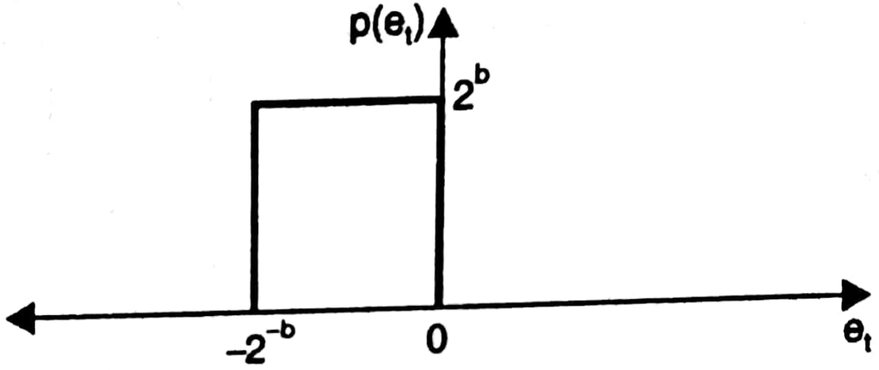
Fixed Point Two's Complement



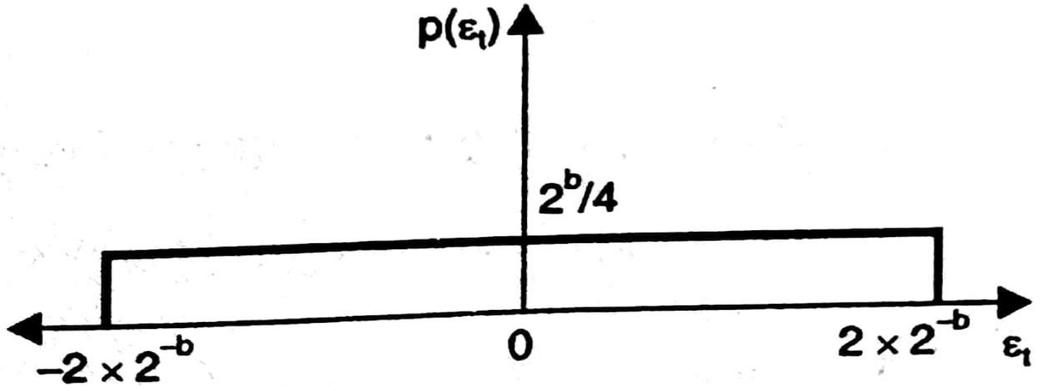
Floating Point Two's Complement



Fixed Point one's Complement



Floating Point one's Complement



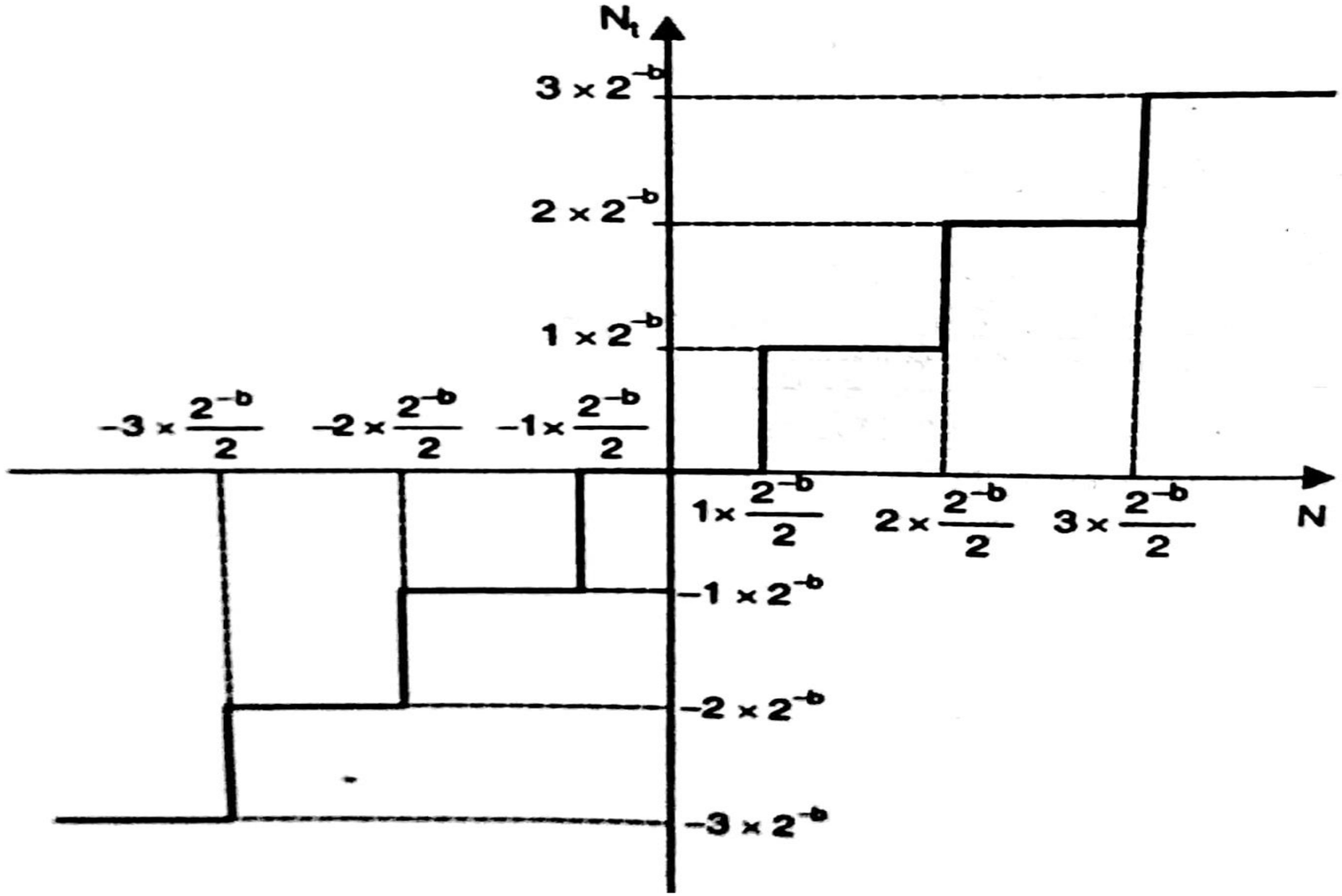
ROUNDING

- Rounding is the process of reducing the size of a binary number to finite word size of b-bits such that the rounded b-bit number is closest to the original unquantized number
- The rounding process consists of truncation and addition
- In rounding of a number of b-bits, first the unquantized number is truncated to b-bits by retaining the most significant b-bits . Then a zero or one is added to the least significant bit of the truncated number depending on the bit that is next to the least significant bit that is retained

ROUNDING

- If the bit next to the least significant bit that is retained is zero then zero is added to the least significant bit of the truncated number
- If the bit next to the least significant bit that is retained is one then one is added to the least significant bit of the truncated number (Here adding one is called rounding up)
- The input-output characteristics of the quantizer used for rounding as shown. The quantization steps are marked on y-axis. The range of unquantized numbers are marked on x-axis

INPUT OUTPUT CHRACTERISTICS OF QUANTIZER USED FOR ROUNDING



INPUT OUTPUT CHARACTERISTICS USED FOR ROUNDING

1. Any positive unquantized number in the range, $\left(1 \times \frac{2^{-b}}{2}\right) \leq N < \left(2 \times \frac{2^{-b}}{2}\right)$, will be assigned the quantization step, 1×2^{-b}.
2. Any positive unquantized number in the range, $\left(2 \times \frac{2^{-b}}{2}\right) \leq N < \left(3 \times \frac{2^{-b}}{2}\right)$, will be assigned the quantization step, 2×2^{-b}, and so on.
3. Any negative unquantized number in the range, $\left(-2 \times \frac{2^{-b}}{2}\right) < N \leq \left(-1 \times \frac{2^{-b}}{2}\right)$, will be assigned the quantization step, -1×2^{-b}.
4. Any negative unquantized number in the range, $\left(-3 \times \frac{2^{-b}}{2}\right) < N \leq \left(-2 \times \frac{2^{-b}}{2}\right)$, will be assigned the quantization step, -2×2^{-b}, and so on.

ROUNDING

- Let N = Unquantized fixed point binary number
- N_r = Fixed point binary number quantized by rounding
- The quantization error in fixed point number due to rounding is defined a

Rounding Error $e_r = N_r - N$

- The range of error due to rounding for all the three formats (i.e., One's complement, Two's Complement and Sign – magnitude) of fixed point representation is same
- In fixed point representation the range of error made by rounding a number of b bits is $-2^{-b}/2 \leq e_r \leq 2^{-b}/2$

ROUNDING

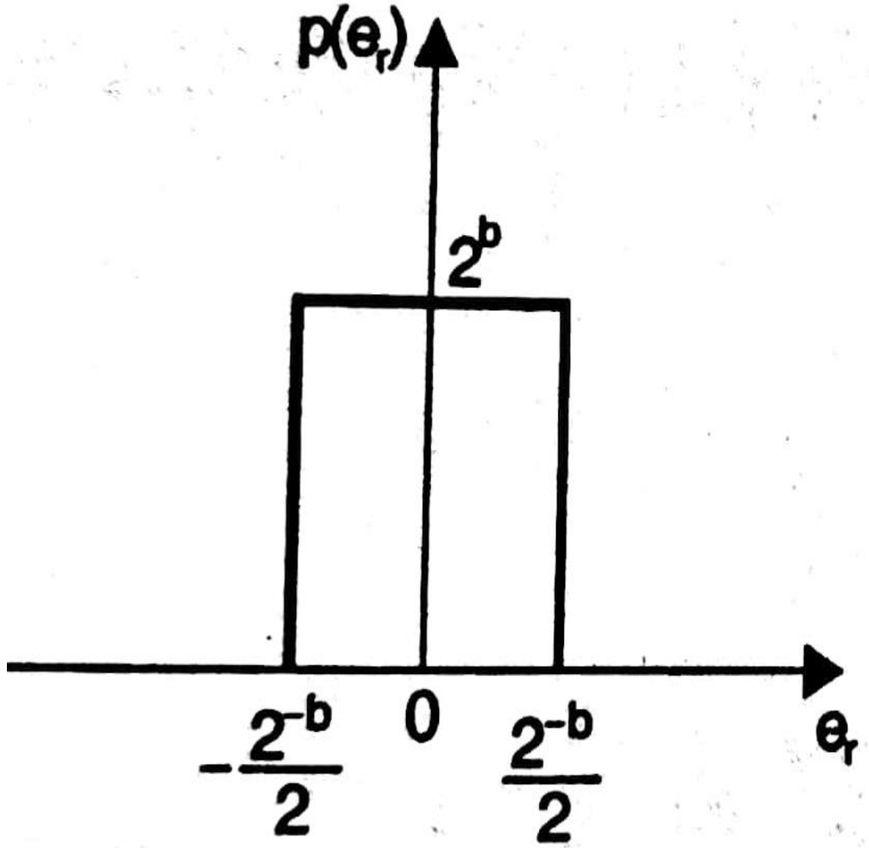
- Let N_f = Unquantized floating point binary number
- N_{rf} = Rounding floating point binary number

$$\mathbf{N_{rf} = N_f + N_f\ \varepsilon_r}$$

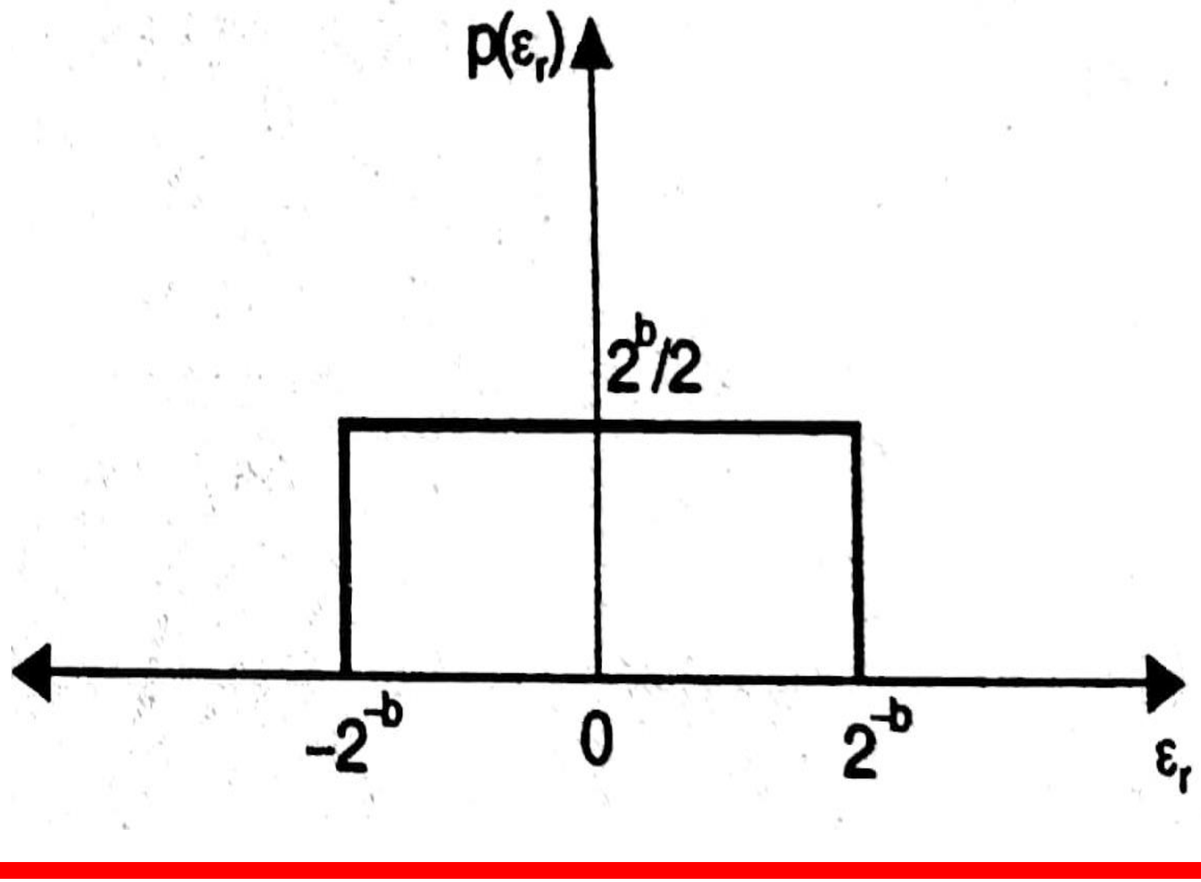
- Where ε_r - relative error due to rounding of a floating point binary number
- Relative error due to rounding is $\boldsymbol{\varepsilon_r = N_{rf} - N_f\ /\ N_f}$
- The range of error due to rounding for all the three formats (i.e., One's complement, Two's Complement and Sign – magnitude) of the mantissa is same

QUANTIZATION NOISE PROBABILITY DENSITY FUNCTIONS FOR ROUNDING

Rounding Fixed Point



Rounding Floating Point



ASSESSMENT

1. The two methods of eliminating these low order bits are --------- & --------
2. Define Truncation.
3. The quantization error in fixed point number due to truncation is defined as ---------
4. Relative error due to truncation is ----------
5. What is meant by rounding?
6. The quantization error in fixed point number due to rounding is defined as ----------
7. Relative error due to rounding is ----------
8. The rounding process consists of ---------- and ----------

THANK YOU